EXECUTIVE SUMMARY

This Background Review Document (BRD) describes the results of a validation study conducted to characterize two *in vitro* basal cytotoxicity tests for determining starting doses for acute oral systemic toxicity assays. The purpose of these tests is to reduce the total number of animals needed for in the *in vivo* tests. As part of this study, methods for two *in vitro* neutral red uptake (NRU) assays using mouse fibroblast (BALB/c) 3T3 cells or normal human epidermal keratinocytes (NHK) were standardized and optimized, the accuracy and validity of the tests were determined using reference chemicals of various toxicities, and computer simulation models were used to estimate the potential reduction in animal usage that could be accomplished by the use of these assays. In addition, high quality *in vivo* lethality and *in vitro* cytotoxicity databases were generated that may be useful in other validation studies for *in vitro* toxicity tests.

The results of the study showed that the 3T3 and NHK NRU test methods are not sufficiently accurate as stand-alone methods to correctly predict acute oral toxicity. However, based on computer simulations for the reference substances tested in this study, the use of these *in vitro* basal cytotoxicity test methods for the selection of starting doses for *in vivo* testing has the potential to reduce both the numbers of animals needed and animal deaths compared to the default procedures.

Introduction and Rationale

Although *in vitro* basal cytotoxicity test methods are not currently regarded as suitable replacements for acute oral systemic toxicity assays (Spielmann et al. 1999; ICCVAM 2001a), such test methods have been evaluated as a means to reduce and refine[2] the use of animals in acute oral systemic toxicity testing. In 1983, an international effort, the Multicentre Evaluation of *In Vitro* Cytotoxicity (MEIC), was initiated to evaluate the relationship of *in vitro* cytotoxicity to acute *in vivo* toxicity. Tests of 50 substances in 61 *in*

---

[2] A reduction alternative is a new or modified test method that reduces the number of animals required. A refinement alternative is a new or modified test method that refines procedures to lessen or eliminate pain or distress in animals or enhances animal well-being (ICCVAM 2003).

vitro assays identified a battery of three human cell line assays that were highly correlated to human lethal blood concentrations. The Registry of Cytotoxicity (RC), a database of 347 substances that currently consists of *in vivo* acute toxicity data from rats and mice and *in vitro* cytotoxicity data from multiple cell lines, was published in 1998 (Halle 1998). A regression formula (the RC millimole regression) constructed from these data was proposed by ZEBET, the German National Center for the Documentation and Evaluation of Alternative Methods to Animal Experiments, as a method to reduce animal use by identifying the most appropriate starting doses for acute oral systemic toxicity tests (Halle 1998; Spielmann et al. 1999). These initiatives (and others) to use *in vitro* cytotoxicity test methods to reduce animal use in acute toxicity testing were evaluated by the International Workshop on *In Vitro* Methods for Assessing Acute Systemic Toxicity in October 2000 ("Workshop 2000"; ICCVAM 2001a). This workshop was organized by the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) and the National Toxicology Program (NTP) Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM).

ICCVAM recommended (ICCVAM 2001a) further evaluation of the use of *in vitro* cytotoxicity data as one of the factors used to estimate starting doses for *in vivo* acute lethality studies based on preliminary information that this approach could reduce the number of animals used in *in vivo* studies (i.e., reduction), minimize the number of animals that receive lethal doses (i.e., refinement), and avoid underestimating hazard. To assist in the adoption and implementation of the ZEBET approach, the *Guidance Document on Using In Vitro Data to Estimate In Vivo Starting Doses for Acute Toxicity* (hereafter referred to as *Guidance Document*; ICCVAM 2001b) was prepared by ICCVAM with the assistance of several workshop participants.

ICCVAM concurred with the Workshop 2000 recommendation that near-term validation studies should focus on two standard basal cytotoxicity assays: one using a human cell system and one using a rodent cell system. Historical data for *in vitro* cytotoxicity testing using 3T3 cells is available through other publications (e.g., Balls et al. 1995; Brantom et al. 1997; Gettings et al. 1991, 1994a, 1994b; Spielmann et al. 1991, 1993, 1996). Historical data for *in vitro* basal cytotoxicity testing using normal human keratinocytes (NHK) cells are also available through other publications (e.g., Gettings et al. 1996; Harbell et al. 1997; Sina et al. 1995; Willshaw et al. 1994).

NICEATM, in partnership with the European Center for the Validation of Alternative Methods (ECVAM), designed a multi-laboratory validation study to evaluate reduction or refinement that might result when using cytotoxicity data from the 3T3 and NHK NRU test methods as part of the weight-of-evidence to estimate starting doses for the Up-and-Down Procedure (UDP; OECD 2001a; EPA 2002a) and the Acute Toxic Class (ATC) method (OECD 2001d). The *Guidance Document* NRU protocols were the initial basis of the NICEATM/ECVAM study protocols. These protocols were derived from examination of the BALB/c 3T3 Cytotoxicity Test, INVITTOX Protocol No. 46 (available at the FRAME-sponsored INVITTOX database [http://embryo.ib.amwaw.edu.pl/invittox/]) and the Borenfreund and Puerner (1985) (3T3 cells) as well as Borenfreund and Puerner (1984) and (Heimann and Rice 1983) (NHK cells). See **Section 2** for a detailed description of the test method protocols.

Test Method Protocol Components

The test method protocol components for the *in vitro* NRU cytotoxicity test methods used in the NICEATM/ECVAM study are very similar for the 3T3 and the NHK cells. The following procedures are common to both cell types:

- preparation of reference substances and positive control
- cell culture environmental conditions
- determination of test substance solubility
- 96-well plate configuration for testing samples
- range finder and definitive testing (48-hour exposure to the reference substance)
- microscopic evaluation of cell cultures for toxicity
- measurement of NRU
- data analysis

The main differences in the test methods are:

- the conditions of propagation of the cells in culture
- the cell growth medium components
- the application of reference substances to the 96-well plate (i.e., different volumes of reference substance solution)

Three testing laboratories participated in testing 72 reference substances, in three phases:

- ECBC: The U.S. Army Edgewood Chemical Biological Center (Edgewood, MD)
- FAL: Fund for the Replacement of Animals in Medical Experiments (FRAME) Alternatives Laboratory (Nottingham, UK)
- IIVS (Gaithersburg, MD)

BioReliance Corporation (Rockville, MD) procured and distributed the coded reference substances and performed solubility tests prior to distribution to the cytotoxicity testing laboratories.

Validation Substances

Reference substances were selected to represent: (1) the complete range of *in vivo* acute oral toxicity (in terms of LD_{50} values where LD_{50} is median lethal dose); (2) the types of substances regulated by various regulatory authorities; and (3) those with human toxicity data and/or human exposure potential. To assure the complete range of toxicity was covered, the Globally Harmonized System of Classification and Labelling of Substances (UN 2005) was used to select 12 substances for each of the five acute oral toxicity categories and 12 unclassified substances. A discussion of characteristics and sources of the reference substances can be found in **Section 3** of the BRD. The set of selected reference substances had the following characteristics:

- 58 of the 72 substances were also included in the RC
- 27 (38%) of the substances had pharmaceutical uses, 15 (21%) had pesticide uses, 8 (11%) had solvent uses, and 5 (7%) had food additive uses. The

remaining substances were used for a variety of manufacturing and consumer products.

- 55 (76%) were organic compounds and 17 (24%) were inorganic compounds; commonly represented classes of organic compounds included heterocyclic compounds, carboxylic acids, and alcohols
- 22 (31%) substances were known or expected to have active metabolites
- many of the selected substances had multiple target organs/effects; including neurological, liver, kidney, and cardiovascular effects

***In Vivo* Rodent Toxicity Reference Data**

Because the *in vitro* NRU cytotoxicity test methods are intended to be used as adjuncts to *in vivo* acute oral systemic toxicity test methods using rats, rodent LD_{50} values from acute oral systemic toxicity tests are the most appropriate reference data for the *in vitro* NRU IC_{50} values (i.e., the concentration of the test substance that reduces cell viability by 50%). *In vivo* LD_{50} reference data for the 72 reference substances were determined from the literature. Limiting the data to studies conducted under Good Laboratory Practice (GLP) guidelines (OECD 1998; EPA 2003a, 2003b; FDA 2003) was not possible since only 3% of the data records were from such studies. While mouse data were considered initially, eventually analyses were restricted to rat data. In total, 485 acute oral LD_{50} values were identified for rats for the 72 reference substances. Reference LD_{50} values for each substance were identified by excluding studies that employed the following materials and methods:

- feral rats
- rats < 4 weeks of age
- anesthetized rats
- test substance administered in food or capsule
- LD_{50} reported as a range or inequality

In vivo reference values were determined, where multiple values existed, by calculating a geometric mean of the values. The reference LD_{50} values for 20 of the 72 substances varied enough from the initial LD_{50} values, which came from the RC database and other summary sources, that the substances were reclassified into different GHS oral toxicity categories.

Test Method Accuracy

Although the 3T3 and NHK NRU test methods are not intended as replacements for acute systemic toxicity assays, the ability of these methods to correctly predict the reference LD_{50} values was used to evaluate their accuracy[3]. The rationale for evaluating the accuracy of LD_{50} predictions was that the animal savings produced by using these *in vitro* test methods to predict starting doses for acute systemic toxicity assays would be greatest when the starting dose is as close as possible to the LD_{50}. An IC_{50}-LD_{50} regression model was used to derive the estimated LD_{50} value using 3T3 or NHK NRU IC_{50} values.

A number of different analyses were done in an attempt to improve the estimation of LD_{50} by the regression. IC_{50}-LD_{50} regressions (millimole units) for each NRU test method and laboratory were developed using the IC_{50} data and reference LD_{50} for the reference substances in the NICEATM/ECVAM validation study. The regressions were not significantly different from a regression for the 58 RC substances (calculated using the RC IC_{50} and LD_{50} data) included among the 72 reference substances (F test; $p = 0.929$ for the 3T3 NRU regression and $p = 0.144$ for the NHK NRU regression).

Discordant substances (i.e., test substances that fit the RC millimole regression poorly) were evaluated. Since the 3T3 and NHK NRU regressions yielded results that were not different from the RC, the RC millimole regression was preferred for analysis of discordant substances since it is based on a larger chemical data set than that used in the NICEATM/ECVAM validation study. Discordant substances from the NICEATM/ECVAM study were analyzed to determine whether there were relationships between their outlier status and physical or chemical characteristics. The lack of fit to the RC millimole regression was correlated with chemical class, boiling point, molecular weight, and log K_{OW}, but not with the insolubility of

---

[3] Accuracy: the agreement between a test method result and an accepted reference value (ICCVAM 2003).

the reference substance in the 3T3 or NHK medium or to the fact that the test method systems had little to no metabolic capability. Since these test methods are based upon basal cytotoxicity, mechanism of toxicity was also considered as a characteristic to explain poor fit to the RC millimole regression. Of the 21 reference substances with specific mechanisms of toxicity that were not expected to be active in the 3T3 and NHK cell cultures, 13 (62%) were outliers (i.e., they fit the RC millimole regression poorly). These substances represented 13/30 (43%) of the outliers for the 3T3 NRU and 13/31 (42%) for the NHK NRU. Information on this analysis is presented in **Section 6.4**.

Additional regressions were developed to improve the RC millimole regression. Substances with *in vivo* LD_{50} values based only on mouse test data were excluded. Substances with mechanisms of toxicity that were not expected to be active in the 3T3 and NHK cell cultures were excluded, leading to the RC rat-only regression excluding substances with specific mechanisms of toxicity. In addition, the RC rat-only data were converted to a weight basis for an additional regression analysis, the RC rat-only weight regression.

Accuracy of the *in vitro* NRU test methods (when used with each of the three IC_{50}-LD_{50} regressions) was characterized by determining the proportion of chemicals for which GHS acute oral toxicity categories were correctly predicted. However, this does not imply that the *in vitro* NRU tests are stand-alone methods that can be used for hazard classification. The accuracy for the prediction of toxicity for substances in the GHS acute oral toxicity categories for $LD_{50} > 2000$ mg/kg was improved by removing substances with specific mechanisms of toxicity from the RC rat-only weight regression (compared with the RC millimole regression). It did not improve the accuracy of category prediction for substances with $LD_{50} < 50$ mg/kg or for substances with $300 < LD_{50} \leq 2000$ mg/kg; however, in the latter case, accuracy was already relatively high. The RC rat-only weight regression excluding substances with specific mechanisms of toxicity improved the overall accuracy for the 3T3 NRU test method from 26% (12/46 test substances) with the RC millimole regression to 46% (21/46 test substances). The RC rat-only weight regression excluding substances with specific mechanisms of toxicity improved the overall accuracy for the NHK NRU test method from 28% (13/47 test substances) for the RC millimole regression to 38%

(18/47 test substances). For each regression evaluated, there was a general trend to underpredict the toxicity of the most toxic chemicals and to overpredict the toxicity of the least toxic chemicals. A detailed discussion of the accuracy analyses is presented in **Section 6.3.**

Test Method Reliability

Intra- and inter-laboratory reproducibility of the 3T3 and NHK NRU IC_{50} data were assessed using analysis of variance (ANOVA), coefficient of variation (CV) analysis, comparison of the laboratory-specific IC_{50}-LD_{50} regressions to one another (for each test method), and laboratory concordance for the GHS acute oral toxicity category predictions. Reproducibility is the consistency of individual test results obtained in a single laboratory (intralaboratory reproducibility) or in different laboratories (interlaboratory reproducibility) using the same protocol and test samples.

Although ANOVA results for the positive control, SLS, IC_{50} for the 3T3 NRU test method indicated there were significant differences among laboratories ($p = 0.006$), a graphical display of the data (see **Figure 7-1**) shows that laboratory means and standard deviations for each study phase overlap one another. Interlaboratory CV values, which ranged from 2% to 10% for the study phases, also indicated that the laboratories were similar. ANOVA results for the SLS IC_{50} for the NHK NRU test method also showed significant differences between laboratories ($p < 0.001$). A different cell culture method at FAL was responsible for SLS IC_{50} differences among the laboratories in Phases Ia and Ib. After harmonization of culture methods with the other laboratories, the laboratory means and standard deviations were quite similar for Phases II and III (see **Figure 7-1**). Interlaboratory CV values for SLS in the NHK NRU test method ranged from 8% (Phase III) to 39% (Phase Ia). Very small slopes ($< |0.001|$) for linear regression analyses of the SLS IC_{50} over time (within each laboratory) for both *in vitro* NRU test methods indicated that the SLS IC_{50} was stable over the 2.5 year duration of the study.

ANOVA results for the reference substances showed significant laboratory differences for 26 substances for the 3T3 NRU test method and seven substances for the NHK NRU test method (see **Table 7-6**). An analysis to determine the relationship, if any, between substance attributes and interlaboratory CV indicated that physical form, solubility, and volatility had little effect on CV. CV seemed to be related, however, to chemical class, GHS acute toxicity category, IC_{50}, and boiling point (see **Section 7.2.2**). Although the ANOVA results and the interlaboratory CV analysis (at least for the 3T3 NRU) seemed to indicate that interlaboratory reproducibility may be less than desired, the comparison of laboratory specific IC_{50}-LD_{50} regressions indicated that the laboratory regressions for both test methods were not significantly different from one another ($p = 0.796$ for the 3T3 NRU and $p = 0.985$ for the NHK NRU). In addition, the laboratory concordance for the prediction of GHS oral toxicity categories ranged from 78 - 85% for the 3T3 NRU and 84 - 91% for the NHK NRU (depending on the regression used). The similarity of the laboratories in LD_{50} predictions (via regression) and GHS toxicity category predictions is considered most significant with respect to the reproducibility analyses since the NRU methods are proposed for use with the regressions in determining starting doses for acute oral toxicity tests.

Animal Welfare Considerations: Reduction, Refinement, and Replacement

For the NICEATM/ECVAM validation study, computer simulation models were used to simulate the UDP and ATC testing of the reference substances tested with the NRU basal cytotoxicity test methods. Reference substances that had only mouse reference LD_{50} data or with known mechanisms of toxicity that were not expected to be active in the 3T3 and NHK cell cultures were not evaluated. The number of animals used for simulated testing and the number of animals that lived or died were determined for the default starting dose and for the NRU-determined starting dose (i.e., one default dose lower than the estimated LD_{50}) with 2000 computer test simulations for each substance and starting dose. The computer simulations accounted for the accuracy of the NRU results with respect to the prediction of LD_{50} values since the accuracy was conferred by the particular regression evaluated

Computer simulation modeling of UDP testing shows that, for the substances tested in this validation study, the prediction of starting doses using the NRU test methods resulted in the use of statistically fewer animals by an average of 1.00 - 1.16 animals (approximately 12%)

when using the RC rat-only weight regression excluding substances with specific mechanisms of toxicity depending upon NRU test method and dose-response slope (of 2 or 8.3). There were no animal savings for chemicals with $50 < LD_{50} \leq 300$ mg/kg when test substances were grouped by GHS toxicity category since animal use was compared with animal used for the default starting dose of 175 mg/kg. However, statistically significant animal savings were as high as 1.75 - 2.22 (19.1 - 20.5%) animals for substances with $2000 < LD_{50} \leq 5000$ mg/kg or $LD_{50} > 5000$ mg/kg. Using the NRU test methods to estimate starting doses also resulted in approximately 0.1 to 0.2 fewer deaths for the simulated UDP testing compared to the default starting dose.

Computer simulation modeling of ATC testing showed that, for the substances tested in this validation study, the prediction of starting doses using the NRU test methods resulted in the use of 1.68 - 1.94 (15.4 - 21.1%) fewer (statistically) animals for the RC rat-only weight regression excluding substances with specific mechanisms of toxicity depending upon NRU test method and dose-response slope (of 2 or 8.3). There were no animal savings for substances with $300 < LD_{50} \leq 2000$ mg/kg when test substances were grouped by GHS toxicity category since animal use was compared with animal use using the default starting dose of 300 mg/kg. Using the RC rat-only weight regression excluding substances with specific mechanisms of toxicity, the highest animal savings for both test methods were for substances with $2000 < LD_{50} \leq 5000$ mg/kg (1.23 [11.0%] - 3.07 [25.8%] animals) and substances with $LD_{50} > 5000$ mg/kg (3.79 [31.8%] - 4.38 [36.5%] animals). Using the NRU IC_{50} values to estimate starting doses for the ATC refined animal use by producing approximately 0.6 to 0.7 fewer animal deaths than when the default starting dose of 300 mg/kg was used.

Practical Considerations

Practical issues to consider for implementation of these cell culture test methods include the need for and availability of specialized equipment, training and expertise requirements, cost considerations, and time expenditure. Good Cell Culture Practice: ECVAM Good Cell Culture Practice Task Force Report 1 (Hartung et al. 2002) encourages the establishment of practices and principles that will reduce uncertainty in the development and application of *in vitro* test methods.

All equipment and supplies are readily available. The NRU test methods are easily transferable to laboratories experienced with mammalian cell culture methods. Much of the training and expertise needed to perform the 3T3 and NHK NRU test methods are common to all mammalian cell culturists. Additional technical training would not be intensive since these test methods are similar in general performance to other *in vitro* mammalian cell culture assays. GLP training should be provided to technicians to ensure proper adherence to protocols and documentation procedures.

Prices for commercial testing for one substance are $1120 to $1850 for *in vitro* NRU cytotoxicity testing to determine the IC_{50} (IIVS, personal communication). It is not clear if the price of an *in vivo* test would be reduced if it were preceded by an *in vitro* cytotoxicity test to set the starting dose. Thus, use of these test methods may not reduce the overall cost of the *in vivo* rat acute oral toxicity test and might increase the cost, but their use can reduce the number of animals needed for a study. Based on cost and technical procedures associated with culture maintenance, the 3T3 cells are less expensive to use and less difficult to maintain than the NHK cells.

Peer Review

ICCVAM has considered the information in this BRD and developed draft recommendations regarding the current uses of these *in vitro* cytotoxicity test methods, and recommendations for future efforts that should be undertaken to advance the usefulness of *in vitro* methods for predicting *in vivo* acute oral toxicity. These draft recommendations are provided in a separate document. As part of the ICCVAM test method evaluation process, an independent international peer review panel will be convened to carry out an independent peer review of the 3T3 and NHK NRU test methods and to comment on the extent that the ICCVAM recommendations are supported by the information and data provided in the BRD. ICCVAM will consider the peer review panel report and public comments, and develop final test method recommendations that will be forwarded to U.S. Federal agencies for their consideration, and where appropriate, incorporation into applicable test guidelines, regulations, and policies.

ICCVAM has also drafted test method performance standards for *in vitro* acute toxicity test methods as a separate document. These proposed standards used the NICEATM/ECVAM validation study results as performance criteria for the future use of *in vitro* test methods to determine starting doses for acute systemic toxicity testing. The test method performance standards may be revised if other methods with better predictability are adequately validated.